第 41 卷　第 1 期
2023 年 3 月

天 文 学 进 展
PROGRESS IN ASTRONOMY

Vol. 41,　No. 1
Mar., 2023

doi: 10.3969/j.issn.1000-8349.2023.01.06

# 面向宇宙再电离探测的基本数据处理方法

何梦钒[1,2,3]，郑　倩[1]，陕欢源[1,3]，郭　铨[1]

(1.中国科学院 上海天文台，上海 200030； 2.中国科学院 国家天文台，北京 100101； 3. 中国科学院大学，北京 100049)

**摘要：** 宇宙再电离时期 (epoch of reionization, EoR) 的探测是 SKA 的重要科学目标之一，也是目前许多 SKA 探路者阵列的首要科学目标。由于宇宙再电离信号非常微弱，因此在数据处理的过程中存在许多难点，如高精度校准、大视场高动态成像等。对默奇森宽场阵列 (Murchison Widefield Array, MWA)、低频阵列 (Low Frequency Array, LOFAR)、21CMA 阵列 (21 Centimeter Array, 21CMA) 等 SKA 低频先导干涉阵列的基本数据处理方法进行了综述，如干扰的识别与去除、数据校准、可视度研究以及成像研究等，并对数据处理时用到的一些常用技术与软件作了相应的介绍与总结。

**关　键　词：** 宇宙学；宇宙再电离；射电干涉阵列；数据处理

**中图分类号：** P159; P161　　　　**文献标识码：** A

## 1　引　言

基于射电干涉综合孔径技术，目前已建成若干射电干涉阵列，如 21CMA 阵列 (21 Centimeter Array, 21CMA[1])、默奇森宽场阵列 (Murchison Widefield Array, MWA[2, 3])、低频阵列 (Low Frequency Array, LOFAR[4])、巨米波射电望远镜 (Giant Meterwave Radio Telescope, GMRT[5])、澳大利亚 SKA 探路者阵列 (Australian Square Kilometre Array Pathfinder, ASKAP[6, 7]) 等。正在建设中的平方公里阵列望远镜 (Square Kilometre Array, SKA) 是多国参与建设的射电望远镜阵列，其覆盖面积将达到一平方公里，孕育着宏伟的科学目标，如宇宙黎明和再电离时期探测[8]、脉冲星搜寻[9]、暂现源研究[10]、宇宙磁场[11]、星系形成[12]等。

**收稿日期：** 2022-02-17；　**修回日期：** 2022-05-31

**资助项目：** SKA 专项 (2020SKA0110100)；国家自然科学基金 (10973069, 10973070)；中科院基础前沿科学研究计划 (ZDBS-LY-7013)；上海市浦江人才计划 (19PJ1410700, 19PJ1410800)

**通讯作者：** 何梦钒，mfhe@bao.ac.cn

射电干涉阵列由许多单天线组成，其数据处理方法与单口径射电望远镜不同。在基于射电干涉阵列的数据处理中，数据校准、大视场高动态成像等方面都还存在着困难与挑战。尤其随着大规模射电望远镜阵列的建成，海量的数据处理也会造成硬件和软件上的压力。解决这些困难，完善数据处理中的各个环节是目前的重要工作，也是实现科学目标的根本保证。

探测宇宙再电离时期是未来 SKA 低频阵列的首要科学目标之一，其观测的 HI 21 cm 信号主要集中在 50 ~ 200 MHz 这一低频范围。由于科技发展和人类活动，在几十到几百 MHz 的低频波段，充斥着无可避免的射电干扰 (radio frequency interference, RFI[13])，在数据处理的过程中进行干扰的识别与去除无法避免。除此之外，前景射电源在低频波段辐射较强，而望远镜的分辨率随观测频率的降低而降低，因此如何在低频射电波段建立较为准确的天空模型，改善校准结果，以及如何扣除低频波段具有较强辐射的前景、微弱再电离信号的识别与提取、高动态范围的大视场成像等都是目前再电离探测数据处理中的难点。除了宇宙黎明和再电离时期探测，射电源低频波段辐射性质的研究、星系团弥散辐射研究、高红移类星体在低频射电波段的辐射研究等都是低频观测的重要科学方向，这些科学目标的实现都依赖于完善的数据处理方法，也同样无法避开 RFI 的扣除、提高校准精度等问题。

`CASA`①, `MIRIAD`[14], `AIPS`②是目前可以使用的面向射电干涉阵列数据处理的基本软件包。SKA 探路者阵列，如 LOFAR, MWA, GMRT, 21CMA 等，目前已经形成了面向不同科学目标的低频射电干涉阵列数据处理的基本流程，并且新的算法在不断更新中。然而，为了应对未来 SKA 更高精度和更大数据量的数据处理需求，目前的数据处理方法在校准、成像以及统计测量等方面依然有待提升。本工作将针对探测宇宙再电离这一 SKA 重要科学目标，对目前 SKA 探路者阵列所涉及的基本数据处理方法以及目前常用的数据处理软件包进行梳理和总结。

第 2 章介绍低频射电干涉阵列观测的基本原理，射电干涉阵列观测数据的基本处理，包括干扰剔除、主瓣改正、校准等；第 3 章介绍宇宙再电离探测的可视度研究和成像研究；第 4 章对面向宇宙再电离探测的基本数据处理流程进行介绍；第 5 章进行总结与讨论。

## 2 低频射电干涉阵列数据处理方法

### 2.1 射电望远镜阵列观测

本小节主要介绍射电干涉阵列的观测数据——可视度 (visibility) 的理论基础，以及干涉阵列的基本参数分辨率、灵敏度的计算方法。

对于两个距离为 $L$ 的天线 $i$ 和 $j$，观测频率为 $\nu$，在时间 $t$ 时，接收到的信号通过天线的电压响应可以表示为：

$$V_i(t,\nu_k) = G_i(t,\nu_k)\int B_i(s)\widetilde{E}_k(s)\mathrm{d}^2 s \tag{1}$$

①`http://casa.nrao.edu`

②`http://www.aips.nrao.edu`

和

$$V_j(t,\nu_k) = G_i(t+\Delta t,\nu_k)\int B_i(s')\widetilde{E}_k(s')\mathrm{d}^2s'\mathrm{e}^{-\mathrm{i}\omega\Delta t} \quad , \tag{2}$$

其中，$G$ 和 $B$ 分别为天线的增益和主瓣响应函数，$\nu_k$ 表示频率，$s$ 表示天空中的位置，$\Delta t$ 表示两个天线接收到同一信号的时间差，$\widetilde{E}_k$ 表示观测信号电场 $E$ 的第 $k$ 阶傅里叶变换。

可视度可以通过两个电压的互相关来计算，即：

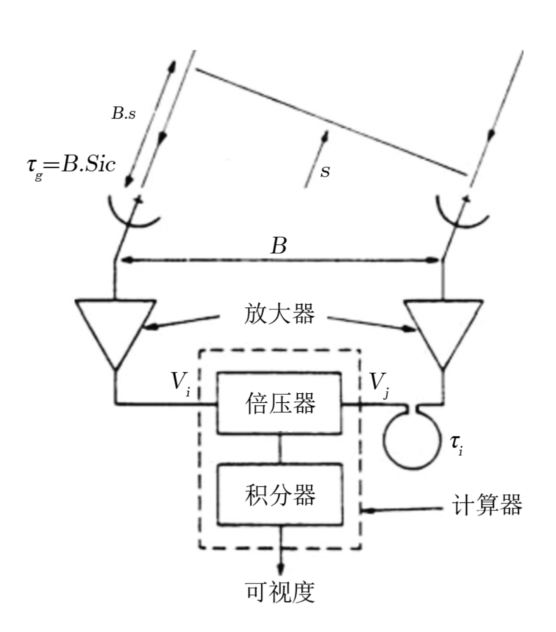
$$V_{ij} = G_iG_j^*\left\langle\left\{\int B_i(s)\widetilde{E}_k(s)\mathrm{d}^2s\right\}\left\{B_j(s')\widetilde{E}_k^*(s')\mathrm{d}^2s'\mathrm{e}^{\mathrm{i}w_k\Delta t}\right\}\right\rangle \quad . \tag{3}$$

如图 1 所示，延迟 $\Delta t$ 与天线之间的距离以及观测目标在天空中的位置有关，可用以波长为单位的地面坐标系 $(u,v,w)$ 以及其对应的天空坐标系 $(l,m,n)$ 表示，从而可得：

$$V_{ij}(u,v,w) = G_iG_j^*\int\int B_{ij}(l,m)|\widetilde{E}_k(l,m)|^2\mathrm{e}^{\mathrm{i}2\pi[ul+vm+w(n-1)]}\frac{\mathrm{d}l\mathrm{d}m}{\sqrt{1-l^2-m^2}} \quad . \tag{4}$$

在小天区近似 $n=\sqrt{1-l^2-m^2}\approx 1$ 下，可得：

$$V_{ij}(u,v) \approx G_iG_j^*\int\int B_{ij}(l,m)|\widetilde{E}_k(l,m)|^2\mathrm{e}^{\mathrm{i}2\pi[ul+vm]}\mathrm{d}l\mathrm{d}m \quad . \tag{5}$$




注：其中 $B$ 为基线长度，$V_i$ 和 $V_j$ 为两个天线输出电压，设备延迟表示为 $\tau_i$，地理延迟表示为 $\tau_g$，由基线长度以及观测目标与天顶之间的夹角决定，对输出的电压两两相关得到可视度信号。

**图 1　低频射电相关干涉原理图**[15]

而射电干涉阵列的空间分辨率可用 $\theta=\dfrac{\lambda}{D}$ 计算，其中 $\lambda$ 为观测信号波长，$D$ 为观测时使用的最长基线的长度。

干涉系统的灵敏度计算方法与单天线的类似，单个天线望远镜的流量密度误差为[15]：

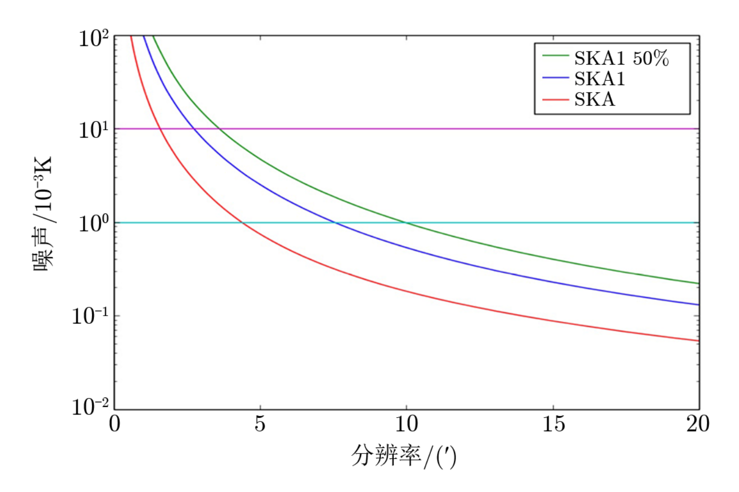
$$\Delta S_v = 2Mk\frac{T_{\mathrm{sys}}\mathrm{e}^{\tau}}{A_{\mathrm{e}}\sqrt{2t\Delta v}} \quad , \tag{6}$$

其中，$M$ 是额外损耗的调整因子，$k$ 为玻尔兹曼常数，$T_{\mathrm{sys}}$ 为系统亮温度，$\tau$ 为大气光深，$A_{\mathrm{e}}$ 为单个天线的有效接收面积，这里的 $t$ 指的是积分时间，$\Delta v$ 为积分的频率带宽。而对于有 $n$ 个独立望远镜的干涉阵列来说，有 $N = n(n-1)/2$ 个相关的天线对，则流量密度的误差为：

$$\Delta S_v = 2Mk\frac{T_{\mathrm{sys}}\mathrm{e}^{\tau}}{A_{\mathrm{e}}\sqrt{2Nt\Delta v}} \quad . \tag{7}$$

所以干涉阵列的灵敏度由观测时间、频率带宽和天线的有效面积等参数共同决定。

对于再电离观测，由于再电离的信号十分微弱，大约只有 0.01 K，灵敏度要达到信号水平需要长时间积分累积，图 2 给出的是红移为 8.95，积分时间 1 000 h 时，观测噪声随图像分辨率的变化[16]；这里图像的积分频率带宽与分辨率对应，即将分辨率换算成对应红移 $z = 8.95$ 处的距离，再把距离换算成对应视线方向的频率范围。可以看出，即使是 1 000 h 的积分时间，对于 SKA1-LOW 也至少需要大约 $4'$，即视线方向上 18.3 kpc 对应的频率带宽的积分，才能使观测噪声达到 EoR 信号以下。




注：红移为 8.95，积分时间等于 1 000 h。不同的曲线代表不同完成度的 SKA 阵列，即其对应公式 (7) 中的 $N$ 不同，SKA 50% (绿色) 指只有 SKA1-LOW 的一半阵列的情况，SKA1 (蓝色) 对应全部 SKA1-LOW，SKA (红色) 则表示所有 SKA2-LOW 阵列的情况，其灵敏度是 SKA1-LOW 的 4 倍。这里测试的 SKA 阵列使用 2012 年发布的 SKA1-LOW 的基线排布方案①。两条横线分别表示 0.01 K (玫红)，和 0.001 K (天空蓝)，0.01 K 为理论估计的 EoR 信号的强度。

**图 2　观测图像噪声随成像分辨率的变化[16]**

此外，在动态范围上，由于再电离信号与前景之间相差 4 ~ 5 个量级，使得观测的动态范围要达到 4 ~ 5 个量级，才能对其进行观测。若要进行再电离时期成像研究，需要在每

①https://www.skatelescope.org/wp-content/uploads/2012/07/SKA-TEL-SKO-DD-001-1_BaselineDesign1.pdf

一个空间/频率分辨率构成的单元内 EoR 信号的信噪比达到 1 以上。现有的设备如 MWA, LOFAR 等阵列，由于灵敏度的限制，可以通过增大分辨率单元，即在高于分辨率的范围内进行平滑，和增加观测时间的做法，以减少噪声，具体可以参考 Zaroubi 等人[17]的做法，他们最终给出的 LOFAR 成像分辨率大约为 20′，而如果 LOFAR 进行功率谱研究，角分辨率大约为 3′。然而，即使是这样的成像分辨率，依然需要 2 400 h 的积分时间才能较好地重构出输入的再电离信号，所以目前国际上已经运行的射电望远镜阵列主要进行再电离信号统计特性，如功率谱的研究。而对于未来的 SKA1-LOW 来说，它的灵敏度足够高，对于 0.001 K 的信号来说，在角分量级的分辨率下，只需要频率积分时间达到 $Bt = 1\,000\,\mathrm{MHz} \cdot h$ 就可以使信噪比达到 1 以上，所以具备成像研究的条件。而在观测视场上，进行再电离时期成像研究需要不小于 1° 的视场以完整呈现再电离气泡。

### 2.2　数据检验

获得射电干涉阵列的观测数据，即可视度数据时，首先可以通过数据的统计性质来检验数据质量，根据观测数据的质量来判断数据是否可以被使用，以及决定后续工作采用的数据处理管线。通常可以通过数据点的数目、平均值及标准差等进行检验。

数据点数目是判断观测数据质量的标准之一，包括不同频率带宽内数据点的数目、不同时间间隔内数据点的数目以及不同基线和极化方向上测量的数据点的数目。在数据处理中，首先对观测频率范围及时间范围内出现的干扰进行剔除，干扰的剔除会按不同基线操作，也会分不同极化方向进行处理。统计干扰剔除前后的数据点数目可用以判断数据质量。

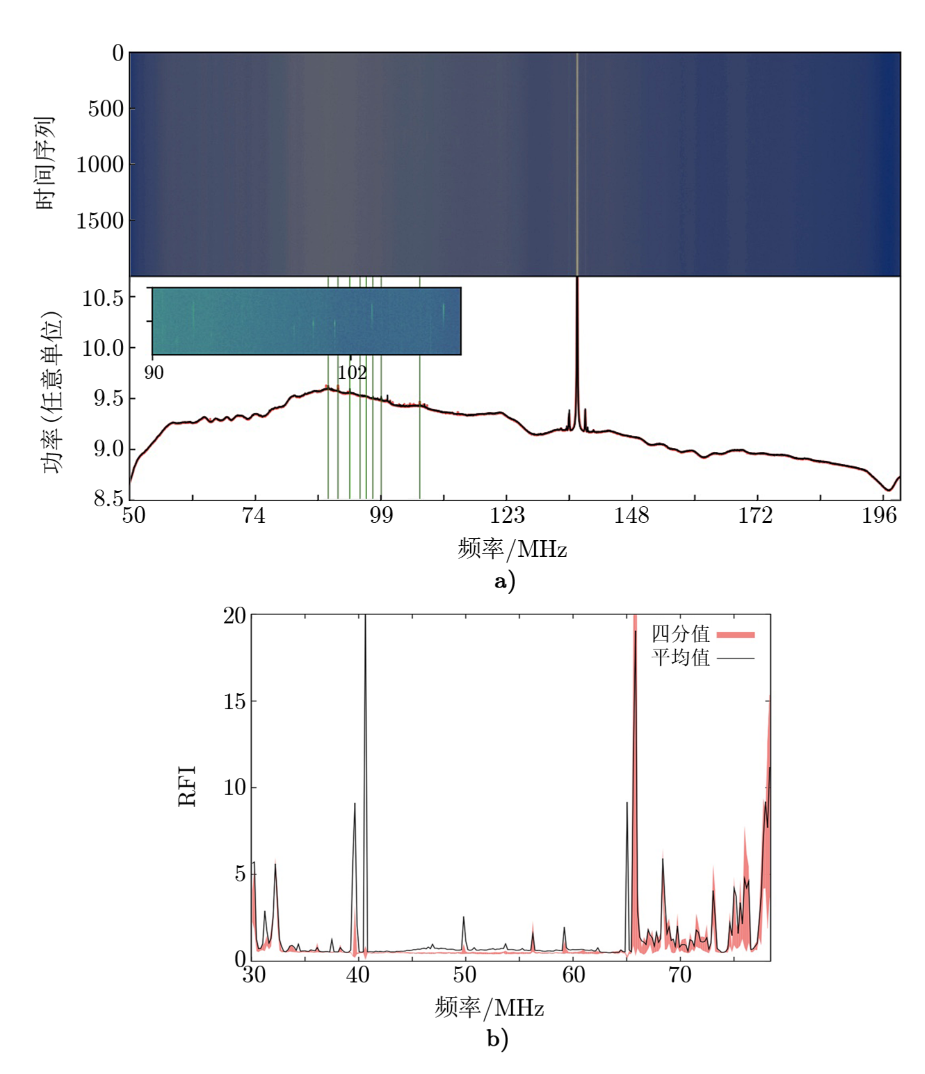
数据点的平均值和标准差是检验观测数据质量的两项重要统计性质，平均值可以估算观测的信噪比并了解天线的增益，而标准差则可用以判断观测数据的可靠性，通过剔除在时间、频率空间和使用不同基线时标准差发生较大变化的数据点，从而减少这些数据点对数据校准造成的影响。

目前的射电阵列数据处理中，有一些常用的工具可以进行数据质量检验，如 `aoqplot`[18] 工具等，MWA 和 LOFAR 等一些低频阵列就是使用 `aoqplot` 进行数据质量的检验。

### 2.3　干扰剔除

在射电观测中，尤其是低频波段，存在射电干扰，需要进行干扰的识别与剔除。虽然射电望远镜阵列一般都会建在射电宁静区域，但是随着科技的发展，理想的射电宁静区域非常稀少，所以干扰信号的识别和剔除是射电望远镜阵列数据处理过程中不可避免的步骤。图 3 给出了 21CMA[19] 和 LOFAR[20] 观测中出现的射电干扰信号。这些干扰主要来自于调频广播、卫星通信、航空通信，以及电离层和对流层引起的散射效应，除此之外还有一些影响时标比较短的干扰信号，如闪电、流星余迹的反射等。

对于不同性质的干扰信号，需要采用不同的方法进行识别和剔除。如特定频率工作的卫星信号，可以在频率空间进行识别和剔除；飞机经过望远镜阵列时散射无线电广播信号，可以在时域进行识别和剔除。对于比较强的干扰信号，无论是在频域还是时域，识别起来相对容易；但对于相对微弱，而且特征不明显的干扰信号，需要通过特定算法进行识别和剔除。同时随着观测数据量的增大，需要通过自动高效的干扰识别方法来保证数据质量，从而满足




注：a) 21CMA 观测信号瀑布图以及 RFI[19]，上半部分是频率信号随时间变化的瀑布图，下半部分黑线是上半部分时间内功率平均后的频谱分布图。绿直线为周边调频广播频段，而 144 MHz 处的强干扰可能来自对讲机信号。b) LOFAR 低频段天线 (low-band antenna, LBA：10 ～ 80 MHZ) 6 h 观测得到的含有 RFI 的频谱[20]，受 RFI 影响较小的子带 (sub band) 中去除掉的数据不到 1%，然而 RFI 影响较大的时需要去除 5% ～ 20% 的数据。

图 3 21CMA[19] 和 LOFAR[20] 观测信号中的 RFI

科学需求。

目前面向低频射电阵列观测数据的干扰识别和剔除方法包括 `AOFlagger`[18] 和 `FLAGCAL`[21] 等。`AOFlagger` 在进行干扰识别时使用了和阈值 (sum threshold) 方法。背景拟合技术以及结合形态算子 (morphological operators) 等，从而能够快速准确地识别射电干涉阵列观测数据中的干扰信号，目前已被应用于 MWA 和 LOFAR 等阵列的数据处理中。而 `FLAGCAL` 基于良好的可视度数据在时间和频率上是连续的原则，来对可视度数据进行识别。然后它通过已知的流量和相位定标源进行校准，再使用球面线性插值到目标场，从而实现对较差的可视度数据进行识别和剔除，其被应用于 GMRT 的数据处理中。在进行干扰的识别和剔除之后，通常会形成一个记录文件，包括被剔除的数据点的相关信息，以便后续查看和调用。

Gao 等人[19] 提出一种对毫秒时间尺度上射频干扰源进行提取的方法，该方法可对输入信号频谱瀑布图，利用 Canny 边缘检测算法与 Hough 寻线算法来识别射电干扰信号，并将这一方法应用到 21CMA 观测数据处理上，发现该算法对于特定的短时标、强干扰信号的证认准确率可以达到 90%。

现有的 RFI 剔除对于一些特定频率 (窄带)、特定宽度的 RFI 能够有效剔除，但仍有一些微弱干扰，无法被识别和剔除。Wilensky 等人[22] 使用 MWA 的快速全息反卷积 (fast holographic deconvolution, FHD) /$\epsilon$ppsilon 管线研究了 RFI 对于 EoR 功率谱探测影响。结果发现，即使是微弱的 RFI 都可能掩盖在 $0.1 < k < 2\ h\cdot \mathrm{Mpc}^{-1}$ 范围内；只有大约 0.01 $\mathrm{K}^2$ 的 EoR 信号，残留 RFI 的总流量值为 1 mJy 的情况下，若真实信号与他们使用的 21cm FAST 生成的乐观模型偏差低于 10% 的话，则 EoR 信号可以在模 $k < 0.9\ h\cdot \mathrm{Mpc}^{-1}$ 的范围内探测到。

**2.4　视场外亮源旁瓣处理**

如果亮源存在于所选定的观测天区周围，其旁瓣会在观测数据中产生一个类似噪声的污染，我们称之为远旁瓣源噪声 (far side-lobe source noise, FSSN)，需要对其进行处理，因此在选取观测天区时，会尽量避免亮源出现在观测视场中以及视场周围。但是，如果基于科学目标的需求，无法避免亮源，那么需要对其进行剔除。剔除观测天区以外亮源的旁瓣影响，一般通过对源进行准确的建模，并且扩大观测天区，从而对其进行旁瓣的改正，保证所需观测的区域不受亮源旁瓣的影响。

如在 21CMA 观测中，21CMA 的观测有效视场是北天极区域 5° 半径的范围，但是有效视场外的亮源产生的旁瓣仍然会影响有效视场。而且 21CMA 阵列采用冗余基线排布，因此亮源的栅瓣 (grating lobes) 也会产生影响。目前采用的处理方法是对较大的视场进行成像，如以北天极为中心，半径 10° 以上的区域，然后在成像的过程中，不但对视场内射电源的旁瓣进行处理，也对视场外的亮源的旁瓣进行处理。

另一方面，FSSN 受观测的 UV 覆盖的影响，随着 UV 覆盖的提高而降低。Mort 等人[23] 使用 SKA-LOW 的分布进行 FSSN 的研究指出，在观测时间超过 6 h 的情况下，FSSN 的水平已经低于热噪声；但是随着观测时间的进一步增加，FSSN 的下降比热噪声下降慢。通过改变阵列的排布或增加阵列的数目提高瞬时孔径覆盖，可以在一定程度上降低 FSSN。

### 2.5 基线、频率以及观测时间选择

如何选择所需的基线、频率以及观测时间的数据点进行下一步的数据处理，取决于科学目标及数据质量。

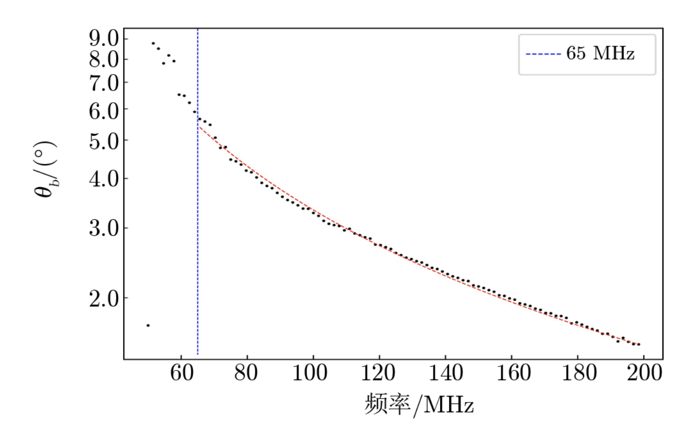
(1) 较短的基线对应较大尺度信息，长基线则对应小尺度信息。长基线的使用可以提高数据的空间分辨率，然而长基线受电离层的影响更为明显，需要对电离层变化时在幅度和相位上造成的影响进行校准和改正。

(2) 数据处理采用的频率带宽与最终的灵敏度相关，为了在数据处理中获得较好的校准效果以及图像质量，需要选择合适的频率带宽。

(3) 在观测时间的选取上，会避免出现较强干扰时的数据，这一步在进行干扰剔除时已经进行了处理。积分时间的长度与灵敏度相关，由式 (7) 可知，$\Delta S_v \propto 1/\sqrt{t}$，积分时间越长，灵敏度越高，为获取较微弱的信号，需要进行长时间的数据累加。但在考虑积分时间的同时，也要考虑无法分辨的弱源产生的致淆极限的影响。

### 2.6 天线主瓣改正

天线主瓣是天线方向图上最大的辐射波束，与天线的方向性有关，并随频率变化。主瓣的宽度定义为场强 $E_k$ 下降至主瓣最大值的 $1/\sqrt{2}$ 时两点之间的宽度，或者功率 $P_k = E_k^2$ 下降到最大值 1/2 时两点之间的宽度，即半功率全宽 (half power beam width, HPBW)。在数据处理中，需要考虑天线阵列主瓣的影响，在不同的观测频率进行改正。精确测量天线主瓣是非常困难的，一些低频射电阵列，如 21CMA 的主瓣可以近似为高斯形式 $F(\theta,\nu) = G(\nu)\mathrm{e}^{-\theta^2/2\theta_\mathrm{b}^2}$，$\theta_\mathrm{b}$ 随频率的变化可以写为：$\theta_\mathrm{b} = 3^\circ.33(\pm0.01)(\nu/100)^{-1.14(\pm0.01)}$，这里频率 $\nu$ 需以 MHz 为单位。Zhao 等人[24]给出 21CMA 的主瓣宽度随频率的变化以及拟合结果，如图 4 所示。




注：$\theta_\mathrm{b}$ 指的是近似为高斯形主瓣高斯分布的标准差，黑色数据点是对子频段带宽为 1.56 MHz 得到的图像上拟合得到的 $\theta_\mathrm{b}$，对频率范围 65 ∼ 200 MHz 的黑色数据点进行拟合得到最佳拟合结果用红色虚线表示，可用表达式 $\theta_\mathrm{b} = 3^\circ.33(\pm0.01)(\nu/100)^{-1.14(\pm0.01)}$ 进行描述，这里频率 $\nu$ 需以 MHz 为单位。

图 4 21CMA 的主瓣随频率的变化[24]

Line 等人[25]对 MWA 的天线主瓣进行了研究，指出天线主瓣随频率的变化带来的不确

定性会对如宇宙再电离探测等科学目标的实现造成不可忽略的影响；而从可视度数据上测量天线主瓣，由于会受到来自仪器、大气以及多种影响因素的共同作用，从而变得十分困难。该工作使用了在对数据进行数字化及相关计算前，从天线接收单元进行主瓣分析等技术。Barry 等人[26]研究了再电离探测对于校准需求后指出，再电离观测要求频率方向上的误差不高于 $10^{-5}$，这也意味着天线主瓣的校准误差造成的信号在频率方向上的起伏须比观测信号低 5 个量级以上，才能进行再电离探测。

### 2.7 校准

实际观测到的射电源的亮度和位置与真实的亮度和位置会出现偏差，因此须对观测数据进行校准之后才具有科学使用价值。校准的方式之一是通过定标源进行校准。定标源可以分为流量和相位定标源。在射电干涉阵列观测时，需要通过观测相位定标源来对相位中心进行校准，同时需要使用流量定标源对最终的观测流量进行校准。在望远镜阵列相位和增益不稳定的情况下，需要在观测的过程中对定标源进行多次观测，通常在观测的开始和结束时进行定标源的观测。如果观测时间较长，需要对定标源进行周期性的观测，从而实现对相位和流量的修正。校准的另一种方式称为自校准，是对天线阵列及仪器本身的校准。可以通过闭合关系来实现天线各基线之间的自校准，从而修正各阵列之间的差异；同时也可以通过使用强信号源，来实现对天线阵列和仪器的校准。

天线真实的指向，即主瓣中心的位置，与预定的指向位置存在偏差。天线指向造成的偏差是方向依赖的，并且会随时间变化。通过观测不同天区的已知位置的射电源，可以得到天线指向误差的方向依赖特性。如果射电源的强度已知，那么可以通过测量射电源的强度与可视度幅度的比值来测量天线在不同方向上的增益。在射电干涉阵列进行观测时，天线的增益如果随频率发生变化，那么就需要对每个频带内的天线增益进行改正。这种频率空间天线增益的改正可以通过选取频谱平滑、辐射较强的定标源来实现。

在对观测的对象进行流量和相位的校准时，地面观测无法避免会受到电离层随时间和频率的变化造成的影响，并且依赖于定标源的校准方法受观测的时间、频率以及观测天区的限制，以及在某些观测中，所需观测的视场还可能缺少理想的定标源，因此校准变得十分困难。除了依赖于定标源的校准方法，在射电干涉阵列数据处理中，通常还会使用自校准和冗余基线校准。如果射电干涉阵列在排布时采用冗余基线，即阵列单元间存在相同的间隔，那么可以采用冗余基线校准来对观测数据进行改正。但由于许多天线阵列并不采用冗余基线的排布模式，因此就无法使用冗余基线进行校准。

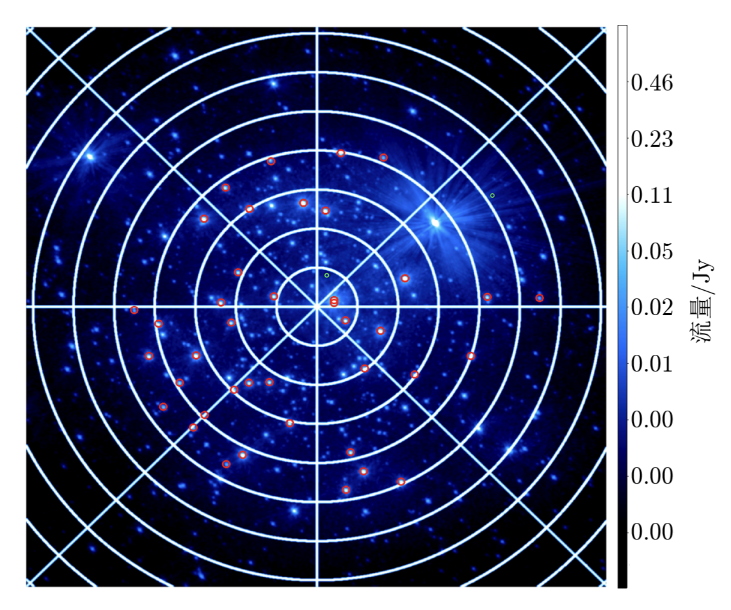
通过输入源的模型来求解天线增益，进行对可视度的改正，基于改正的可视度形成新的模型，再次进行对天线增益的求解。如此循环，直到形成模型趋近于输入模型。通过这种方法可以对天线系统间的差异进行校准。`SAGECAL` (Space Alternating Generalized Expectation Maximization Calibration)① 是可以实现这一步骤的射电阵列校准软件包之一[27]。`SAGECAL` 使用的天空模型可以包含 `shapelets` 拟合的成分、盘状结构、环状结构、高斯成分以及标准的点源。

①https://github.com/nlesc-dirac/sagecal

MWA 射电望远镜阵列目前采用冗余基线校准和基于天空模型的校准方法[28]。由于 21CMA 阵列也采用冗余基线的排布形式，因此在校准方法上也可以采用冗余基线的校准方法和基于天空模型的校准方法。

### 2.8 天空模型

Barry 等人[26]使用 MWA EoR 观测 FHD/$\epsilon$ppsilon 管线进行仪器校准影响的研究表明，使用一个理想的校准模型，可以完美地重构出输入的 EoR 信号，但是现有的星表都无法涵盖所有的源，而已有的源中也无法做到测量的流量、位置、形态毫无误差。另一方面，前景扣除或是前景规避也一定程度上依赖于人们对前景的认知。所以，建立理想的天空模型可以提高校准精度，同时也是研究射电源性质以及宇宙黎明和再电离时期探测前景去除的重要工作。射电源天空模型的建立首先是射电源的搜索与识别，然后是模型的建立，尤其是对结构复杂的弥散源进行模型建立。以 21CMA 数据处理为例，使用了 40 个射电源形成天空模型，对北天极区域半径 5° 范围内的天区进行校准，如 Zhao 等人[24]给出的图 5 所示。




注：21CMA 观测的北天极 (North Celestial Pole, NCP) 附近，每个圆环为半径 1° 的天空图像，最中心为 NCP 90°。红色圆圈圈出的是校准时使用的 40 个定标源。

图 5 21CMA 频率 50 ~ 200 MHz 的生成的天空图像[24]

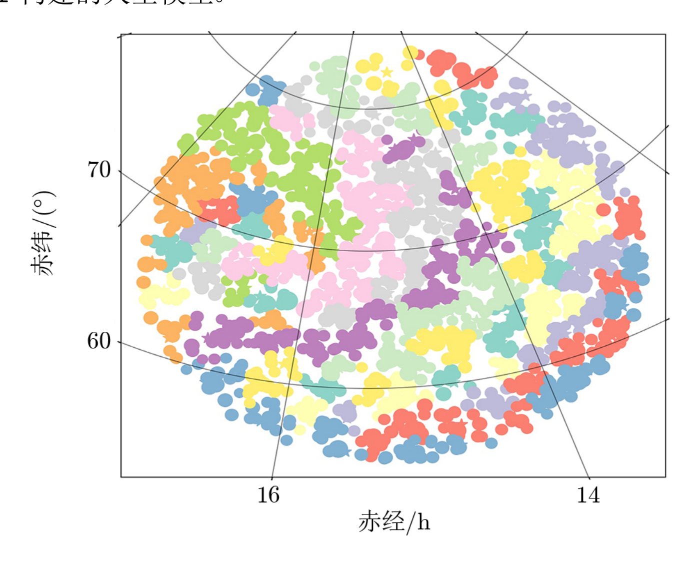
目前有许多工具都可以用来进行射电源搜索和建模，主要有五种。

(1) PyBDSF (Python Blob Detection and Source Finder)① (原为 PyBDSM) 是基于图像的射电源搜索 Python 软件包[29]，可以输出射电源列表文件，包含源的位置、流量等信息。对于结构复杂的射电源，PyBDSF 将使用多个椭圆高斯的形式进行拟合；对于延展性较强的源，PyBDSF 也可以对其进行 Shapelets 分解。由于射电图像会存在许多处理过程中引入的假结

①https://www.astron.nl/citt/pybdsf/

构，这些假结构也会被 PyBDSF 识别成射电源，可以通过对流量限制等参数进行调整，尽量避免识别出假源。

(2) 基于搜索出的射电源，LSMTool[①] 软件包可进一步选择并构建天空模型。除了包含 PyBDSF 搜索的源，LSMTool 还可以通过特定的分类函数对 CLEAN 形成的成分进行射电源天空模型的建立。对于弥散源，其弥散成分可以通过 Shapelets 来进行模型建立，从而对包含大尺度弥散结构的射电源，如星系团射电晕、射电遗迹等建立较为精确的模型。图 6 展示了使用 LSMTool 构建的天空模型。




注：不同的颜色代表不同的天空区块 (sky patches)，不同的形状表示源的不同形态类型，如圆形表示点源，星型表示高斯形态的源等。

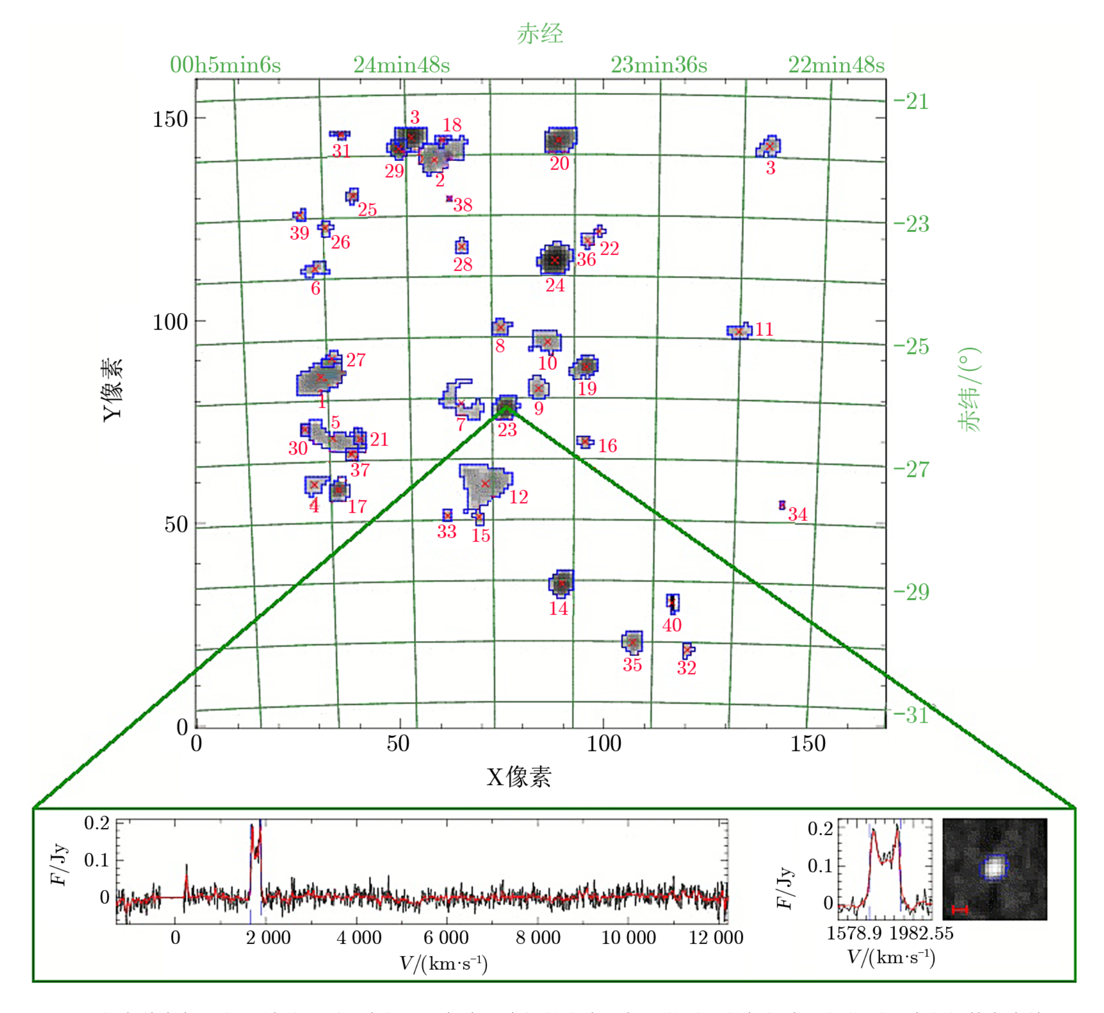
图 6　使用 LSMTool 构建出的天空模型的简单图示

(3) Aegean 也是一种基于射电图像的源搜索识别工具[30, 31]，它主要针对的是高斯形态的致密源的查找与拟合。与 PyBDSF 类似，Aegean 对于图像进行源查找拟合之后输出包含源的位置、流量等信息的列表文件，并且对于延展源以及解析的结构，Aegean 采用多椭圆高斯叠加的拟合模型。Aegean 程序主要可分为两个部分。第一部分，“岛”的搜索与界定。这里“岛”指的是在设定信噪比阈值以上的像素连成一片的区域，它由两个可由用户设定的信噪比阈值决定：一个是种子阈值，只有像素点上的信噪比大于这个阈值，才认为有源存在；以这些像素点为中心，向外延伸的区域中，信噪比大于另一个洪阈值的区域被认为是这个源存在影响的区域。只有这些区域内的数据后续才被用来进行源的位置、流量、形态等信息的测量。第二部分，根据原图计算曲率图来获得拟合采取的组分数目和每个组分初始参数，进而对证认出的岛进行参数拟合。

(4) Duchamp[32] 可用于三维数据中源的查找，它同样可以应用于二维数据，甚至是一维数据上。为了提高源查找的完备性和准确性，程序提供了几种预处理方法以减少数据

[①] https://github.com/darafferty/LSMTool

中的误差，根据预处理的方法不同又可以分为 Duchamp basic, Duchamp smooth, Duchamp àtrous。其中 Duchamp smooth 是对数据使用定义的核进行卷积平滑，而 Duchamp àtrous 是一种利用基于多分辨率小波变换的 àtrous 算法来降低数据误差的方法。Duchamp 对预处理过后的数据进行源的搜索，搜索方式与 Aegean 类似，也是采用设定信噪比阈值的方法，识别出源所在的 pixel，对这些识别出来的源像素进行参数化，输出积分流量、加权中心、主轴等信息的表格，同时输出的还有源的图像 (见图 7)。



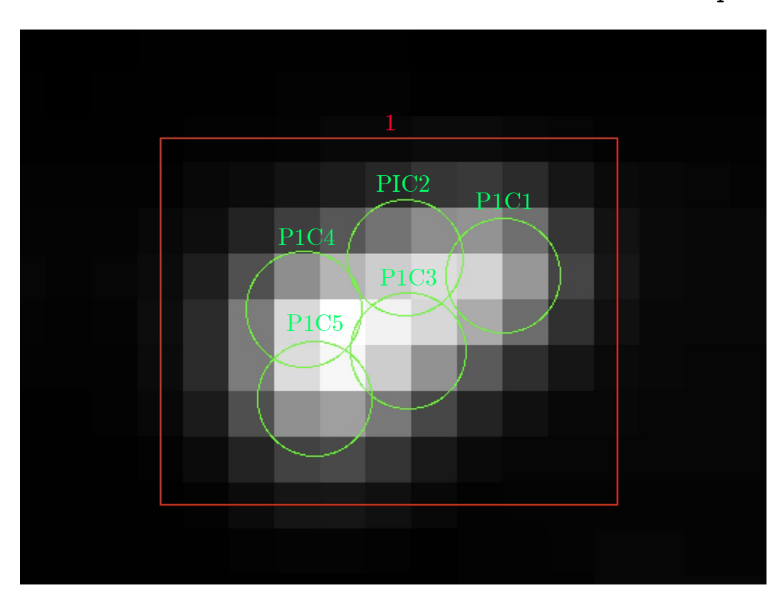

注：绿色线条框出的区域是识别出来的源，每个源中间的十字叉标注的是源的加权中心的位置，边上的数字为输出列表中源的序号。

图 7 Duchamp 源查找结果①

(5) Buildsky[33] 是基于不损失精度的前提下，对源使用最少的组分而开发的，被用于 LOFAR 数据的自动化天空建模。Buildsky 通过对给定的源选取合适的自由度以达到建立

①https://www.atnf.csiro.au/people/Matthew.Whiting/Duchamp/

最简模型的目的。例如点源只有一个自由度，即它的形状，更复杂的源自由度更高，最佳的自由度数目根据信息论标准来决定。Buildsky 对一个较为复杂延展源的致密中心进行拟合的结果如图 8 所示。值得注意的是，LOFAR 对于延展的结构会使用 Shapelets 对其建模。




注：绿色圆圈为对这一部分进行建模使用的 5 个点源。

图 8　Buildsky 建模结果[①]

除此以外的射电源搜索与识别的软件还包括 IFCA[34, 35], APEX[36]，以及可以进行实时暂现源搜索的 PySE[4] 等。

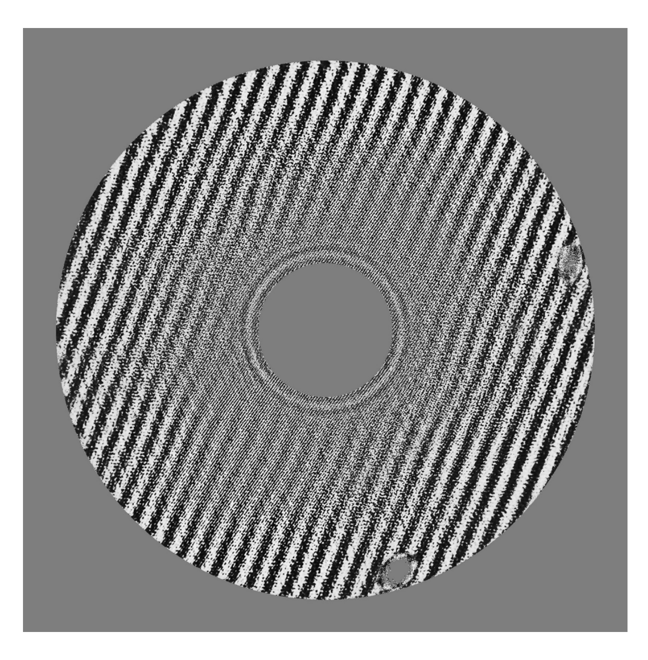
## 3　宇宙再电离探测的可视度研究和成像研究

### 3.1　可视度研究

低频射电阵列可以在 UV 空间直接进行可视度研究，如在宇宙再电离探测中，可以在 UV 空间直接构建功率谱，对宇宙再电离信号的统计性质进行研究，从而避免在成像过程中由傅里叶变换以及格点化带来的误差。来自于天空和仪器的噪声通常可以近似为高斯形式，直接的傅里叶变换也不会影响噪声的性质。然而，如果采用一些非线性的方法，如 CLEAN 方法对图像进行处理，那么噪声的性质就会偏离原有的高斯属性。因此，直接对可视度进行研究可以较容易地对噪声进行处理。图 9 展示了 21CMA 阵列 24 h 观测得到的可视度数据，可视度数据构成的圆环中，由内到外对应频率 50 ~ 200 MHz，每个频率对应的圆上，不同位置对应不同时间的观测，24 h 的观测数据正好构成一个完整的圆形。

由于电离层的影响和天线增益的变化，需要对可视度的实部和虚部进行校准。通常情况

[①]https://support.astron.nl/LOFARImagingCookbook/shapelets.html



注：其中环形由内到外对应频率 50 ～ 200 MHz，每个频率对应的圆上，不同位置对应不同时间的观测，24 h 的观测数据正好构成一个完整的圆形，圆环在右边偏上，和底部偏右有两处缺失是对数据在 UV 平面初步进行筛查时会直接剪切掉的存在干扰，或缺少观测的数据。

**图 9 21CMA 观测的 UV 图像**

下，对可视度的校准可以通过自校准来实现。在自校准效果不理想的情况下，可以通过闭合关系对可视度的实部和虚部进行校准。然后通过在 UV 空间绘制观测数据点，或者绘制观测数据点幅度和观测尺度 (基线长度) 的关系来进行数据检查。

针对不同的科学目标，在数据处理的过程中选择不同的权重方式。

(1) 均匀权重。每个格点的权重 $W$ 反比于该格点上所包含数据点的个数 $N$，即：

$$W_{\rm u}(u_i, v_j) = \frac{1}{N(u_i, v_j)} \quad , \tag{8}$$

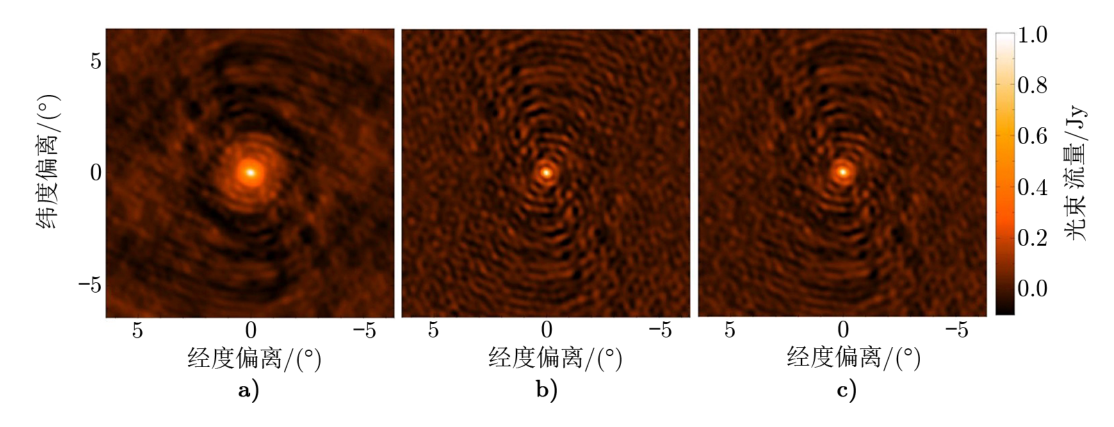
其中，$u_i$, $v_j$ 表示 UV 空间上的格点位置。均匀权重有利于实现较高的分辨率，并且旁瓣较弱，但是在一定程度上牺牲了灵敏度。

(2) 自然权重。每个格点的权重为该格点上数据点的数目，假设格点上的数据点数目服从正态分布，则：

$$W_{\rm n}(u_i, v_j) = N(u_i, v_j) \approx \frac{1}{\sigma^2(u_i, v_j)} \quad . \tag{9}$$

自然权重得到的图像具有最低的噪声水平，但图像的分辨率较差，旁瓣明显，当 UV 覆盖均匀时，自然权重与均匀权重的效果相似。

(3) Briggs 权重 (Robust 权重)。兼顾分辨率和旁瓣影响，它通过调节稳健参数来控制前面所述两种方法的作用程度。图 10 给出了几种权重的成像效果。




注：a) 自然权重；b) 均匀权重；c) Briggs 权重。

图 10　几种权重作用效果的比较①

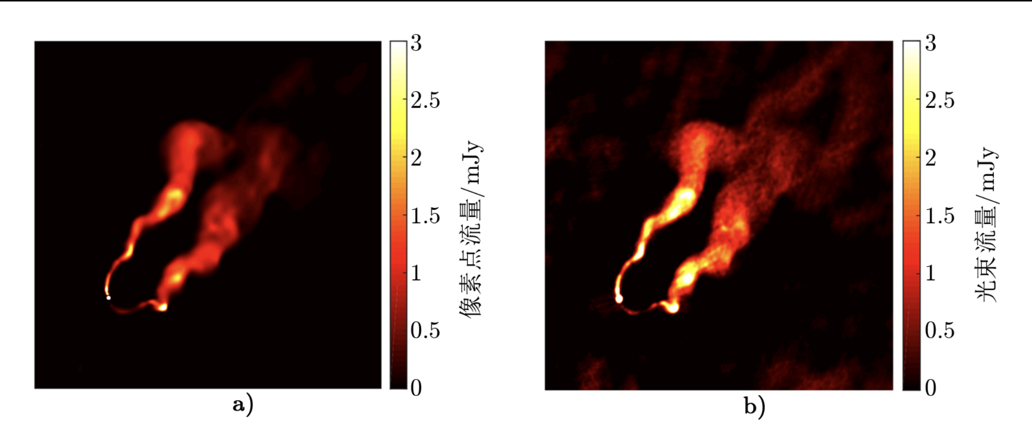
### 3.2　成像研究

除了对 UV 空间的数据进行统计研究，还可以依据科学目标的需求，通过对可视度数据进行逆傅里叶变换，从而生成图像。图 5 是图 9 经过逆傅里叶变换生成的天图。在使用逆傅里叶变换时，需要对数据进行格点化。由于望远镜阵列 UV 覆盖的不完备，需要通过 CLEAN 等算法进行修正，形成理想的射电图像，Cotton-Schwab CLEAN[37] 是比较常用的 CLEAN 算法之一。

在使用射电干涉阵列观测数据进行成像时，需要考虑 W 项 (UVW 坐标中，W 项表示视线方向) 的影响，可以使用的方法主要包括 W 项投影 (W-projection)[38−40] 和 W 项叠加 (W-stacking)[40]。21CMA 阵列分东西和南北两条基线，如果只使用东西基线进行观测，就可以简化数据处理过程，避免 W 项的影响。MWA, LOFAR 以及未来 SKA 阵列的天线阵排布则无法避免考虑 W 项的影响。在目前 MWA, LOFAR 等阵列的数据处理中，可以使用 WSClean 算法，WSClean 采用 W 项叠加的算法[40]，在进行 CLEAN 时考虑了 W 项的影响，这也是基于射电干涉阵列进行大视场成像的需求。PURIFY 也是基于射电干涉阵列数据重建图像的方法[41]，使用 W 项投影和 W 项叠加的方法进行大视场成像。Pratley 等人[41] 认为 PURIFY 获得的图像比 CLEAN 算法获得的图像具有更大的动态范围。成像动态范围对不同的源来说有所区别，Pratley 等人[41] 使用 PURIFY 对来自甚大天线阵 (Very Large Array, VLA) 和澳大利亚致密天线阵 (Australia Telescope Compact Array, ATCA) 的数据进行的测试显示，PURIFY 的成像动态范围在 $7\times10^5\sim1.1\times10^6$ 之间，而其对应的 CLEAN 方法成像的动态范围只有数百 (见文献 [41] 中的表 3)。其测试的源均为具有复杂结构的源，都能得到较好的成像效果，并且其残差图像的均方根要小于使用 CLEAN 得到的结果。Pratley 等人将 3C129 射电星系使用 CLEAN 的成像与 PURIFY 成像进行了对比，结果如图 11 所示。

在进行格点化时，如果在成像时使用 W 项投影的方法，那么需要考虑 W 项投影平面的

① https://science.nrao.edu/science/meetings/2016/15th-synthesis-imaging-workshop/documents/wilner_vla16.pdf




注：a) PURIFY；b) CLEAN。PURIFY 比 CLEAN 具有更大的成像范围，有些情况下甚至能提高一个量级，得到的重构结果也包含更少的污染。

图 11 PURIFY 和 CLEAN 重构出来的 3C129 射电星系[41]



数量，同时还要考虑主瓣和空间分辨率的影响。对于图像空间的格点化，在数据量比较大时需要考虑格点化算法的计算成本。WSClean 中使用了图像域格点化 (image-domain gridder, IDG) 的方法进行格点化，IDG 可以同时进行 W 项投影和 A 项投影，且计算成本较低。这里的 A 项投影是对方向依赖影响进行改正的算法[42]。Van der Tol 等人[43]开发了通过 AW 投影，在对图像进行格点化时，对方向依赖性影响进行改正的方法，该方法在计算效率和精度上与传统的 W 项投影方法相当，同时因为加入了 A 项投影，可以对快速变换的方向依赖的影响进行改正。Offringa 等人[44]讨论了格点化在探测宇宙再电离信号功率谱测量中的影响，该工作认为格点化不会对功率谱测量造成影响，图像空间进行格点化的结果更为理想，也适用于在大视场成像时使用 W 项叠加的情况。与进行可视度研究相似，在进行图像处理的过程中同样需要考虑不同的权重，如均匀权重、自然权重以及 robust 权重，从而在保证灵敏度的情况下得到所需尺度的信息。

目前，射电干涉阵列成像如何提高成像精度、实现高动态范围成像、大视场成像等依然是亟待解决的问题。

## 4 面向宇宙再电离探测的基本数据处理流程

本章主要对 SKA 低频先导干涉阵列，如 MWA, LOFAR 等阵列的面向再电离探测数据处理流程进行介绍。对 MWA, LOFAR 和 GMRT 在数据处理中常用的软件进行了总结，如表 1 所示。

LOFAR 基本的数据处理流程如下：(1) 数据检查；(2) 干扰剔除 (AOFlagger)；(3) 天线增益校准、相位校准、视场外亮源剔除等；(4) 基于 SAGECAL 的校准；(5) 基于 WSClean 的成像；(6) 射电源搜索及天空模型建立 (PyBDSF, LSMTool)。在具体的数据处理中，会根据

表 1　各个干涉阵列数据处理使用的方法或软件

| 阵列 | RFI 去除 | 校准 | 成像 | 天空模型 |
|---|---|---|---|---|
| MWA | AOFlagger[18] | RTS[45, 46], FHD[47] | WSClean[40] | Aegean[31] |
| LOFAR | AOFlagger | Sagecal-CO[27] | WSClean | buildsky[33] |
| GMRT | SVD，FLAGCAL[21] | SPAM(AIPS)[48, 49] | SPAM(AIPS) | PyBDSM |

注：表格列举的是各阵列进行再电离探测功率谱探测时使用的方法或软件，也是各阵列目前使用的比较常见的工具，在具体的数据处理中，根据观测科学目标，数据情况等，有些研究者可能会选择采用其他的方法或软件进行数据处理。

观测和数据情况以及科学目标进行调整。如图 12 展示的是 LOFAR 测量 EoR 功率谱上限的时候使用的数据处理流程，观测数据首先进入预处理流程，包含使用 AOFlagger 进行 RFI 的标记，以及频率方向上的平均。然后使用 Sagecal 通过北天极包含 28 755 个定标源的天空模型进行校准。校准时可分为两部分，方向无关的校准 (DI) 和方向相关的校准 (DD)。方向无关的校准主要不考虑方向依赖，在较大的天区内对增益、相位等进行大致校准；方向相关的校准则是将数据按照天区进行更加精细的划分，如图 12 将观测天区分成 122 份，对每个小天区，分别进行增益、相位等校准。MWA 中这两部分的校准都可以在 RTS 中进行，对于 GMRT 则包含在 SPAM 处理管线中。使用 WSClean 对校准过的数据进行成像后，便将测量值转化为亮温度表示，然后使用高斯过程回归进行前景去除[50]，最后进行功率谱的测量。

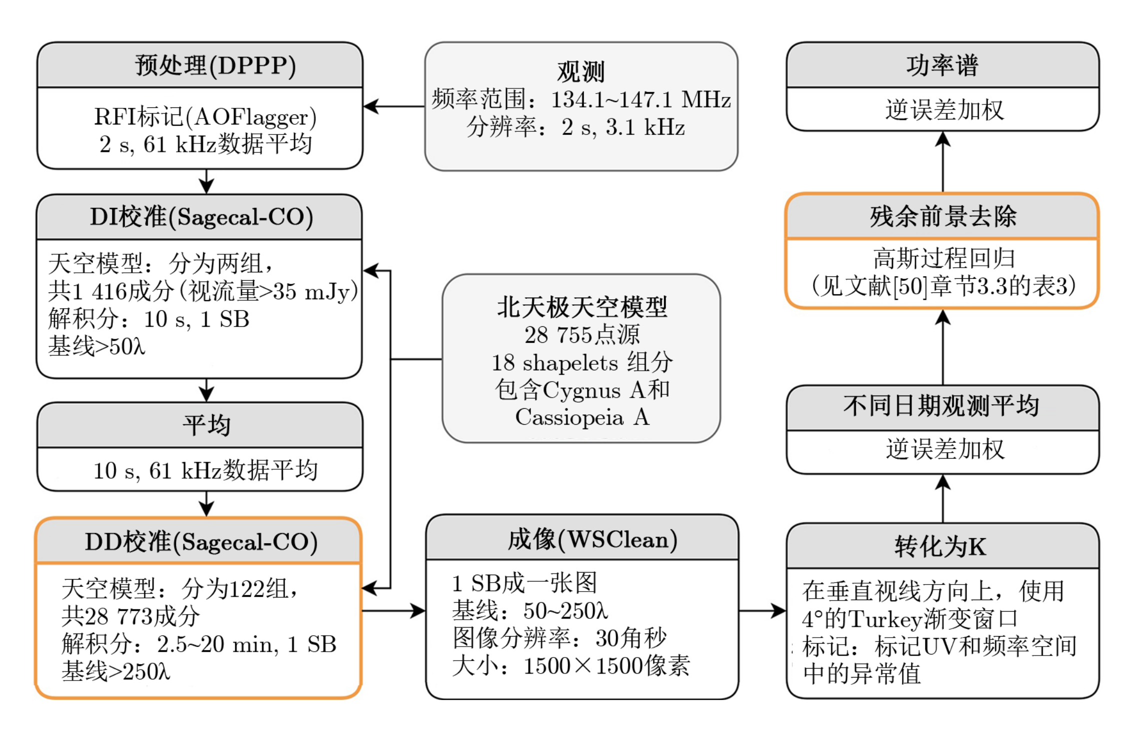



图 12　LOFAR 高频段天线 (120 ~ 240 MHZ) EoR 功率谱测量数据处理流程图[50]

MWA 进行 EoR 功率谱测量的管线主要包括：实时系统 (real time system, RTS) /宇宙学 HI 功率谱测量管线 (cosmological HI power spectrum estimator, CHIPS[51]) 以及 FHD/$\epsilon$ppsilon 管线，其中 RTS 和 FHD 主要用于校准和成像，CHIPS 和 $\epsilon$ppsilon 用于功率谱的测量。RTS 包含了可视度积分、循环校准以及成像三个主要部分，其中循环校准部分能够利用全频率的方向相关的波束响应、电离层的建模与校正、使用包含点源和延展源的天空模型进行校准和源剥离等手段，实现对 MWA 数据进行高程度的校准。RTS 更为详细的流程可以参考文献 [3] 中的图 10。FHD 是 MWA 的另一套可以对测量的可视度数据生成校准过后天空图像的管线。它最初是设计用来进行高效率反卷积的程序，它利用全息映射函数，在无精度损失的情况下，提高全息反卷积的速度；但是现在已经成为 EoR 功率谱分析中进行校准和成像的重要一步。它包含三个核心功能：(1) 生成天空模型的可视度数据，用以校准和前景扣除；(2) 对校准过后的数据进行格点化；(3) 进行成像。

而功率谱测量管线中，CHIPS 是不需要对数据进行成像，可以直接从校准过后的可视度数据中测量功率谱及误差的管线。它采用最大似然估计，在充分考虑仪器和前景残留对于测量数据的协方差的影响情况下，最大程度提取宇宙学信息。$\epsilon$ppsilon 则是从三维图像数据上直接测量功率谱，它能够在积分图像上计算数据、模型、残差的功率，进行观测噪声的测量，测量误差的估计等，帮助人们获得可靠的功率谱上限。Barry 等人[52] 在使用 MWA 观测数据测量 EoR 功率谱时，比较了来自 RTS/CHIPS 管线和 FHD/$\epsilon$ppsilon 管线的结果，两条管线给出的功率谱上限具有较好的一致性，$\Delta^2$ 随 $k$ 模具有一致的变化，RTS/CHIPS 管线能够恢复出更多的 $k$ 模，而 FHD/$\epsilon$ppsilon 管线则在小 $k$ 模处具有更低的功率谱上限。总体上看，MWA 的数据处理流程与 LOFAR 相似，包含相似的流程，且也使用 `AOFlagger`, `WSClean` 等进行射电干扰的去除和成像。在进行巡天数据的大视场成像时，会对经过校准以及 `WSClean` 后的图像进行马赛克图像拼接。

GMRT 的 RFI 识别主要使用单值分解方法 (singular value decomposition, SVD) 和 `FLAGCAL` 包 (在 2.3 节中有介绍)，SVD 方法主要用于宽带干扰的剔除。SVD 剔除干扰，是指利用单值分解的方法将数据分解为不同的模，因为干扰与信号特征的不同，会对应在不同的模上，将干扰对应的模扣去，从而剔除干扰。这种方法也可以应用于平滑前景的扣除，如 FASTICA 就是基于这种方法。

GMRT 也有一条功能比较齐全的数据处理管线——源剥离与大气建模 (Source Peeling and Atmospheric Modeling, SPAM[48, 49]) 管线，它是基于 AIPS 的 `Python` 模块，主要包含预处理和主处理两部分。其中预处理，输入 LAT 格式的可视度数据，以及记录望远镜观测情况的 FLAG 文件，使用流量定标源进行增益和带宽校准，进行初步的 RFI 标记，然后将初步校准的可视度数据拆分为不同 TGSS 观测指向的数据，输入到主处理流程。主处理流程包含精细的 RFI 标记、方向无关的校准、方向相关的校准、电离层色散效应的建模与改正、宽场成像、星表的提取等，将预处理输入的数据转化为最终观测图像。更加详细的处理流程可以参考文献 [49] 的图 A.1, A.2, A.3。由于前面提到过的，诸如 RFI、仪器效应、电离层改正、完善天空模型等原因，目前这些 EoR 数据处理方法能达到的功率谱上限依然高于理论预计的 EoR 信号功率。

## 5 总　结

随着射电望远镜阵列的建设，射电数据处理方法的改进成为实现科学目标的重要前提。本工作对面向宇宙再电离探测的低频射电干涉阵列数据处理的基本方法进行了概述。由于篇幅的限制，并没有涉及到所有的细节问题。而且针对不同科学目标的观测需求，数据处理的技术也会有很大差异，需要针对不同的科学目标选择数据处理的方案以及开发新的算法。如 CASA, MIRIAD, AIPS 等面向射电干涉阵列数据处理的软件包可以满足基本的数据处理需求，但随着对高动态高精度成像、海量高效的数据处理、高精度校准等要求的提高，这些软件包都具有局限性。基于目前已经在运行中的射电干涉阵列 21CMA, LOFAR, MWA, GMRT 等，数据处理的技术和方法在不断地发展中。尤其在面向未来 SKA 大型射电望远镜阵列的海量数据处理时，如何兼顾精度与效率，对数据处理的各个环节都提出了更高的要求。

数据校准是射电干涉阵列数据处理中最大的难点之一。如果对原始的非格点化的可视度数据进行校准，那么将造成计算资源的巨大消耗；同时为了获取更为理想的校准结果，还要兼顾时间和频率分辨率，在计算、存储方面都是巨大的挑战。然而数据校准却是实现某些科学目标，如宇宙再电离探测的统计测量，无法避免的数据处理需求。针对未来 SKA 阵列的海量数据处理需求，将采用区域中心的建设方案，实现原始观测数据到科学用户的对接，将在区域中心上搭建统一的前期数据处理平台，缓解数据分发到科学用户过程中的压力。

除了本工作讨论的干扰识别与去除、数据校准、成像等基本的数据处理步骤，针对不同低频射电阵列以及观测模式，还会涉及到如多波束成像、巡天观测模式时的马赛克图像拼接等。另外，本工作的讨论主要集中在软件方面。在面向未来大型射电阵列的数据处理中，还要结合软件和硬件的使用来获得理想的数据处理结果，如在计算过程中合理使用现场可编程逻辑门阵列 (field programmable gate array, FPGA)、图形处理器 (graphics processing unit, GPU) 等。



**参考文献：**

[1] Zheng Q, Wu X P, Johnston-Hollitt M, et al. ApJ, 2016, 832(2): 190
[2] Bowman J D, Cairns I, Kaplan D L, et al. PASA, 2013, 30: e031
[3] Tingay S J, Goeke R, Bowman J D, et al. PASA, 2013, 30: e007
[4] van Haarlem M P, Wise M W, Gunst A W, et al. A&A, 2013, 556: A2
[5] Paciga G, Chang T C, Gupta Y, et al. MNRAS, 2011, 413(2): 1174
[6] Hotan A W, Bunton J D, Harvey-Smith L, et al. PASA, 2014, 31: e041
[7] Hotan A W, Bunton J D, Chippendale A P, et al. PASA, 2021, 38: e009
[8] Koopmans L, Pritchard J, Mellema G, et al. arXiv e-prints, 2015: arXiv:1505.07568
[9] Kramer M, Stappers B. arXiv e-prints, 2015: arXiv:1507.04423
[10] Fender R, Stewart A, Macquart J P, et al. arXiv e-prints, 2015: arXiv:1507.00729
[11] Johnston-Hollitt M, Govoni F, Beck R, et al. arXiv e-prints, 2015: arXiv:1506.00808
[12] Prandoni I, Seymour N. arXiv e-prints, 2015: arXiv:1412.6512

[13] Sokolowski M, Wayth R B, Lewis M. arXiv e-prints, 2016: arXiv:1610.04696
[14] Sault R J, Teuben P J, Wright M C H. Astronomical Data Analysis Software and Systems IV, 1995, 77: 433
[15] Wilson T L, Rohlfs K, Hüttemeister S. Tools of Radio Astronomy, 2013
[16] Mellema G, Koopmans L, Shukla H, et al. arXiv e-prints, 2015: arXiv:1501.04203
[17] Zaroubi S, de Bruyn A G, Harker G, et al. MNRAS, 2012, 425(4): 2964
[18] Offringa A R, Wayth R B, Hurley-Walker N, et al. PASA, 2015, 32: e008
[19] 高文帅, 赵碧轩, 郭铨, 等. 天文学进展, 2022, 40(2): 284
[20] Offringa A R, DE Bruyn A G2011 XXXth URSI General Assembly and Scientific Symposium, 2011: 1
[21] Prasad J, Chengalur J. Experimental Astronomy, 2012, 33(1): 157
[22] Wilensky M J, Barry N, Morales M F, et al. MNRAS, 2020, 498(1): 265
[23] Mort B, Dulwich F, Razavi-Ghods N, et al. MNRAS, 2017, 465(3): 3680
[24] Zhao B X, Zheng Q, Shan H Y, et al. Research in Astronomy and Astrophysics, 2022, 22(1): 015012
[25] Line J L B, McKinley B, Rasti J, et al. PASA, 2018, 35: e045
[26] Barry N, Hazelton B, Sullivan I, et al. MNRAS, 2016, 461(3): 3135
[27] Yatawatta S, Zaroubi S, de Bruyn G, et al. 2009 IEEE 13th Digital Signal Processing Workshop and 5th IEEE Signal Processing Education Workshop, 2009: 150
[28] Li W, Pober J C, Hazelton B J, et al. ApJ, 2018, 863(2): 170
[29] Mohan N, Rafferty D. PyBDSF: Python Blob Detection and Source Finder, 2015. http://ascl.net/1502.007
[30] Hancock P J, Murphy T, Gaensler B M, et al. MNRAS, 2012, 422(2): 1812
[31] Hancock P J, Trott C M, Hurley-Walker N. PASA, 2018, 35: e011
[32] Whiting M T. MNRAS, 2012, 421(4): 3242
[33] Yatawatta S, de Bruyn A G, Brentjens M A, et al. A&A, 2013, 550: A136
[34] López-Caniego M, Herranz D, González-Nuevo J, et al. MNRAS, 2006, 370(4): 2047
[35] López-Caniego M, Vielva P. MNRAS, 2012, 421(3): 2139
[36] Makovoz D, Marleau F R. PASP, 2005, 117(836): 1113
[37] Schwab F R. AJ, 1984, 89: 1076
[38] Cornwell T J, Golap K, Bhatnagar S. IEEE Journal of Selected Topics in Signal Processing, 2008, 2(5): 647
[39] Tasse C, van der Tol S, van Zwieten J, et al. A&A, 2013, 553: A105
[40] Offringa A R, McKinley B, Hurley-Walker N, et al. MNRAS, 2014, 444(1): 606
[41] Pratley L, McEwen J D, d'Avezac M, et al. MNRAS, 2018, 473(1): 1038
[42] Bhatnagar S, Cornwell T J, Golap K, et al. A&A, 2008, 487(1): 419
[43] van der Tol S, Veenboer B, Offringa A R. A&A, 2018, 616: A27
[44] Offringa A R, Mertens F, van der Tol S, et al. A&A, 2019, 631: A12
[45] Mitchell D A, Greenhill L J, Wayth R B, et al. Real-Time Calibration of the Murchison Widefield Array. 2008, 2: 707
[46] Riding J L, Mitchell D A, Webster R L. Astronomical Data Analysis Software and Systems XXV, 2017, 512: 257
[47] Sullivan I S, Morales M F, Hazelton B J, et al. ApJ, 2012, 759(1): 17
[48] Intema H T, van der Tol S, Cotton W D, et al. A&A, 2009, 501(3): 1185
[49] Intema H T, Jagannathan P, Mooley K P, et al. A&A, 2017, 598: A78
[50] Mertens F G, Mevius M, Koopmans L V E, et al. MNRAS, 2020, 493(2): 1662
[51] Trott C M, Pindor B, Procopio P, et al. ApJ, 2016, 818(2): 139
[52] Barry N, Wilensky M, Trott C M, et al. ApJ, 2019, 884(1): 1

# The Overview of Data Processing for the Detection of the Epoch of Reionization

HE Meng-fan[1,2,3], ZHENG Qian[1], SHAN Huan-yuan[1,3], GUO Quan[1]

*(1. Shanghai Astronomical Observatory, Chinese Academy of Sciences, Shanghai 200030, China; 2. National Astronomical Observatories, Chinese Academy of Sciences, Beijing 100012, China; 3. University of Chinese Academy of Sciences, Beijing 100049, China)*

**Abstract:** Data processing plays a crucial role in achieving the scientific objectives based on radio interferometers. Detection of the Epoch of Reionization (EoR) is one of the key science projects for current low frequency radio interferometers and the coming Square Kilometre Array (SKA). There are still many difficulties and challenges in data processing of radio interferometers. With the construction of SKA, higher sensitivity, resolution and massive amount of data pose new challenges for data processing in both softwares and hardwares. High precision calibration, wide field and high dynamic imaging are required in the EoR detections. This work summarizes the basic data processing methods of low-frequency radio interferometers, such as radio frequency interference identification and removal, calibration, studies on visibilities and images, etc. This work also introduces some common techniques and softwares used by current SKA Pathfinders, such as MWA, 21CMA, LOFAR, GMRT and the EoR power spectra estimate pipeline used in SKA Pathfinders.

**Key words:** cosmology; epoch of reionization; interferometers; data processing